ANOVA

Alan T. Arnholt

Last compiled: May 10, 2022 at 09:37:59 AM

Contents

Chapter 25 Notes

Objectives:

- Overview of ANOVA
- Mathematics of ANOVA
- Assumptions of ANOVA

The idea behind analysis of variance is to find out where the variance in the data lives. We have two possible estimates for the standard deviation of the errors. One is based on the standard deviation of the data differences from the group means, and one is based on the differences of the group means from the overall mean. We compare the differences between the means of the groups ($SS_{\text{Between = Treatment}}$) with the variation within the groups ($SS_{\text{Error = Residual = Within}}$).

When the differences between means are large compared to the variation within the groups, we reject the null hypothesis and conclude the means are not all equal.

The first step in analysis of variance problems is to make a comparative boxplot.

We write our model as $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, where μ is the overall mean, τ_i are the treatment effects, and ε_{ij} are the errors.

If the null hypothesis of all group means are equal, which is equivalent to $H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0$, is true, the differences of the group means from the overall mean will be small compared to the differences of the data from the group means.

To estimate each of the parameters in our model, we use what you would expect:

Parameter	Estimate	Name
μ	$\bar{\bar{y}}$	Grand Mean
τ_i	$\bar{y}_i - \bar{\bar{y}}$	Treatment Effect(k of these)
ε_{ij}	$y_{ij} - \bar{y}_i$	Residual

Consider the identity

$$y_{ij} - \bar{\bar{y}} = (\bar{y}_i - \bar{\bar{y}}) + (y_{ij} - \bar{y}_i) \tag{1}$$

which partitions the deviation of any observation from the grand mean into two parts. The first part, $(\bar{y}_i - \bar{\bar{y}})$, is the deviation of the i^{th} treatment mean from the grand mean. The second part is the deviation of the observation from the i^{th} treatment mean.

If all of the n_i values are the same, we have a balanced design... which makes the calculation easier.

Squaring and summing both sides of (1) produces

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(\bar{y}_i - \bar{\bar{y}}) + (y_{ij} - \bar{y}_i)\right]^2 \\ &= \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{\bar{y}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \\ &\quad + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{\bar{y}})(y_{ij} - \bar{y}_i) \qquad (2)\end{aligned}$$

However, the cross product in (2) is zero since

$$\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i) = \sum_{j=1}^{n_i} y_{ij} - n_i\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} - n_i \cdot \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} = 0.$$

After lots of algebra, we can derive the expression

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{\bar{y}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \qquad (3)$$

which says the total variability in the data can be partitioned into two parts. The quantity $\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{\bar{y}})^2$ measures the difference between the observed treatment means and the grand mean. Specifically, it is a measure of variability due to the treatments and is denoted $SS_{\text{Treatment}}$ (sum of squares due to treatments). The quantity $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$ measures the differences of observations within a treatment from the treatment mean, which must be due to error and is referred to as SS_{Error} (sum of squares due to error). The quantity on the left-hand side of the equals sign in (3) is called the total sum of squares corrected for the mean and is denoted SS_{Total}. The symbolic representation of (3) is

$$SS_{\text{Total}} = SS_{\text{Treatment}} + SS_{\text{Error}} \qquad (4)$$

Since there are a total of $\sum_{i=1}^{k} n_i = N$ observations, SS_{Total} has $N-1$ degrees of freedom. One degree of freedom is lost for estimating μ with the grand mean, $\bar{\bar{y}}$.

$SS_{\text{Treatment}}$ has $k-1$ degrees of freedom since there are k treatment means and SS_{Error} has $N-k$ degrees of freedom.

To adjust for the number of treatments, $SS_{\text{Treatment}}$ is divided by its degrees of freedom, $k-1$. The resulting quantity is known as the **mean square treatment** $MS_{\text{Treatment}} = \frac{SS_{\text{Treatment}}}{df_{\text{Treatment}}}$ and is also called the **between treatments error variance**.

In order to know whether the $MS_{\text{Treatment}}$ value is large, it is compared to an estimate of σ^2, namely, MS_{Error}, where $MS_{\text{Error}} = \frac{SS_{Error}}{df_{Error}}$, which is also called the **within treatments error variance.**

Note that MS_{Error} can be expressed as

$$\hat{\sigma}^2 = MS_{\text{Error}} = \frac{SS_{Error}}{df_{Error}} = \frac{\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2\right]}{df_{\text{Error}}} \qquad (5)$$

If the term within the square braces is divided by its degrees of freedom $(n_i - 1)$, it is easy to recognize that quantity as the sample variance for the i^{th} treatment:

$$S_i^2 = \sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_i)^2}{n_i - 1},\ i = 1, 2, \ldots, a \tag{6}$$

Combining the sample variances, a single estimate of the population variance emerges as

$$\begin{aligned}\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \cdots + (n_k-1)S_a^2}{(n_1-1)+(n_2-1)+\cdots+(n_k-1)} &= \frac{\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2\right]}{\sum_{i=1}^{k}(n_i-1)} \\ &= \frac{SS_{\text{Error}}}{N-k} = MS_{\text{Error}}\end{aligned}$$

The pooled estimate of the variance from the two-sample t-test has now been generalized for k different samples.

If there are no differences among the k treatment means, $MS_{\text{Treatment}}$ is an unbiased estimate of σ^2, and the ratio of $MS_{\text{Treatment}}/MS_{\text{Error}}$ will be close to 1. If differences actually exist among the k treatment means, then the ratio, $MS_{\text{Treatment}}/MS_{\text{Error}}$ should be larger than 1. In fact, it can be shown that

$$E(MS_{\text{Error}}) = \sigma^2 \quad \text{and} \quad E(MS_{\text{Treatment}}) = \sigma^2 + \sum_{i=1}^{k} \frac{n_i \tau_i^2}{k-1}$$

implying that when H_0 is false, $E(MS_{\text{Treatment}}) > E(MS_{\text{Error}})$ since some $\tau_i \neq 0$. When H_0 is true, $\tau_i = 0$ for all i and $E(MS_{\text{Treatment}}) = E(MS_{\text{Error}}) = \sigma^2$. With a little effort, it can be shown that

$$\frac{MS_{\text{Error}}}{\sigma^2} \sim \frac{\chi^2_{df_{\text{Error}}}}{df_{\text{Error}}} = \frac{\chi^2_{N-k}}{N-k}$$

regardless of whether H_0 is true or not, and that

$$\frac{MS_{\text{Treatment}}}{\sigma^2} \sim \frac{\chi^2_{df_{\text{Treatment}}}}{df_{\text{Treatment}}} = \frac{\chi^2_{k-1}}{k-1}$$

when H_0 is true independently of MS_{Error}.

When H_0 is true, MSE and MST are similar, so the ratio is close to 1. If H_0 is false, MS_T will be larger than MS_E.

The sampling distribution of $MS_{\text{Treatment}}/MS_{\text{Error}} \sim F_{k-1;\,N-k}$.

Thus, H_0 is rejected in an α-level test if $F_{\text{obs}} > f_{1-\alpha;\,k-1,\,N-k}$, where $F_{\text{obs}} = MS_{\text{Treatment}}/MS_{\text{Error}}$.

In general, big F-statistics usually imply small p-values.

Assumptions To do an analysis of variance successfully, the assumptions must be met.

1. Independence — ✓ Randomization
    - Surveys — each group is representative

- Experiments — subjects randomly assigned to groups and/or treatments

2. Equal variance of treatment groups. Check similar spread of side-by-side boxplots of residuals $y_{ij} - \bar{y}_i$ If the spread changes, consider $\log y = y_\star$ in experiments, $\sqrt{y} = y_\star$ in observational studies
3. Normal Population — check for outliers in any group

Example

Of the 23 first year male students at State U. admitted from Jim Thorpe High School, 8 were offered baseball scholarships and 7 were offered football scholarships. The University admissions committee looked at the students composite ACT scores (shown in the table), wondering if the University was lowering their standards for athletes. Assuming that this group of students is representative of all admitted students, what do you think?

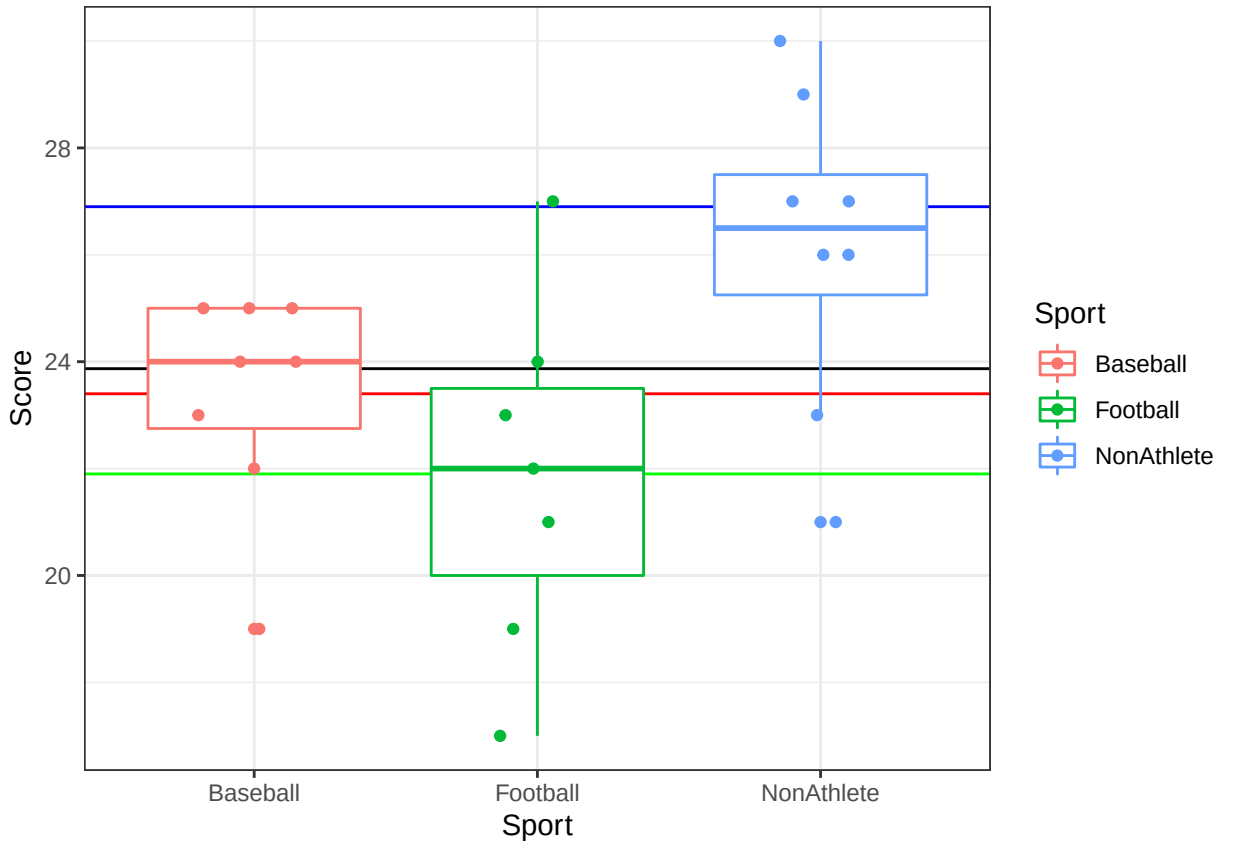
```
Baseball <- c(25, 22, 19, 25, 24, 25, 24, 23)
NonAthlete <- c(21, 27, 29, 26, 30, 27, 26, 23)
Football <- c(22, 21, 24, 27, 19, 23, 17)
DF <- data.frame(Score = c(Baseball, NonAthlete, Football),
                 Sport = c(rep("Baseball", 8), rep("NonAthlete", 8),
                           rep("Football", 7)))
GM <- mean(DF$Score)
GM
```

```
[1] 23.86957
```

```
results <- DF %>%
  group_by(Sport) %>%
  summarize(mean(Score), sd(Score))
results
```

```
# A tibble: 3 x 3
  Sport      `mean(Score)` `sd(Score)`
  <fct>              <dbl>       <dbl>
1 Baseball            23.4        2.07
2 Football            21.9        3.29
3 NonAthlete          26.1        2.95
```

```
# Examine the data
ggplot(data = DF, aes(x = Sport, y = Score, color = Sport)) +
  geom_hline(yintercept = c(23.86957, 23.4, 21.9, 26.9),
             color = c("black", "red", "green", "blue")) +
  geom_boxplot() +
  geom_jitter(width = 0.2, height = 0) +
  theme_bw()
```




```
SSTreat <- 8*(23.375 - GM)^2 + 7*(21.85714 - GM)^2 + 8*(26.125 - GM)^2
SSTreat
```

```
[1] 71.00163
```

```
MSTreat <- SSTreat/2
MSTreat
```

```
[1] 35.50082
```

```
SSError <- sum((Baseball - 23.375)^2 + (Football - 21.857143)^2 + (NonAthlete - 26.125)^2)
SSError
```

```
[1] 155.6276
```

```
MSError <- SSError/20
MSError
```

```
[1] 7.781378
```

```
(F <- MSTreat/MSError)
```

```
[1] 4.562279
```

```
(pvalue <- pf(F, 2, 20, lower = FALSE))
```

```
[1] 0.02331881
```

```
#
anova(lm(Score ~ Sport, data = DF))
```

```
Analysis of Variance Table
```

```
Response: Score
          Df  Sum Sq Mean Sq F value  Pr(>F)
Sport      2  71.002  35.501  4.5629 0.02331 *
Residuals 20 155.607   7.780
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is evidence all the group means are not equal.

Treat df = # treatments -1

Total df = # treatments $\times$ # repetitions -1.

Treat df + Error df = Total df, so Error df = Total df $-$ Treat df.

In general the relationship between the tstat in a regular t-test and the F statistic is that $t^2_{\text{stat}} = F_{\text{stat}}$.

Objectives:

- Confidence interval and test for single comparison
- CI for multiple comparisons
- What can go wrong

To test $H_0 : \mu_1 - \mu_2 = 0$, we need:

$SE(\bar{y}_1 - \bar{y}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where $s_p = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}{N-k}}$, the pooled standard error from all the groups.

The t-test statistic is $\frac{\bar{y}_1 - \bar{y}_2}{SE(\bar{y}_1 - \bar{y}_2)}$, which will have $N - k$ degrees of freedom and a t distribution.

If we wish to do multiple comparisons, we should use a critical value of t equal to $t_{N-k;1-\alpha/2\omega}$ where ω is the number of pairs in the Bonferroni method. Other multiple comparison methods include Tukey, Dunn-Sidak, and Scheffe. http://sci2s.ugr.es/keel/pdf/specific/articulo/35723.pdf has a good explanation of several of the different multiple comparisons.

The $(1-\alpha)\cdot 100\%$ Bonferroni confidence interval for $\mu_1 - \mu_2$ is $\bar{y}_1 - \bar{y}_2 \pm ME$ where the margin of error, $ME = t_{N-k;1-\alpha/2\omega}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. Typically, you will evaluate computer output: if zero is in the CI, there is no evidence of a difference.

A Tukey CI is similar with $ME = q(1-\alpha, r, df_{\text{Error}})\cdot\sqrt{\frac{MS_{\text{Error}}}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ where $q(1-\alpha, r, df_{\text{Error}})$ depends on the number of treatments and total number of data points, not on the individual treatments, so it's the same for all rows in any given experiment.

The command in R to look up values from a q distribution is `qtukey(q, nmeans = , df = dferror)`. Where `q` is a vector of quantiles (area to the left... $1-\alpha_e$).

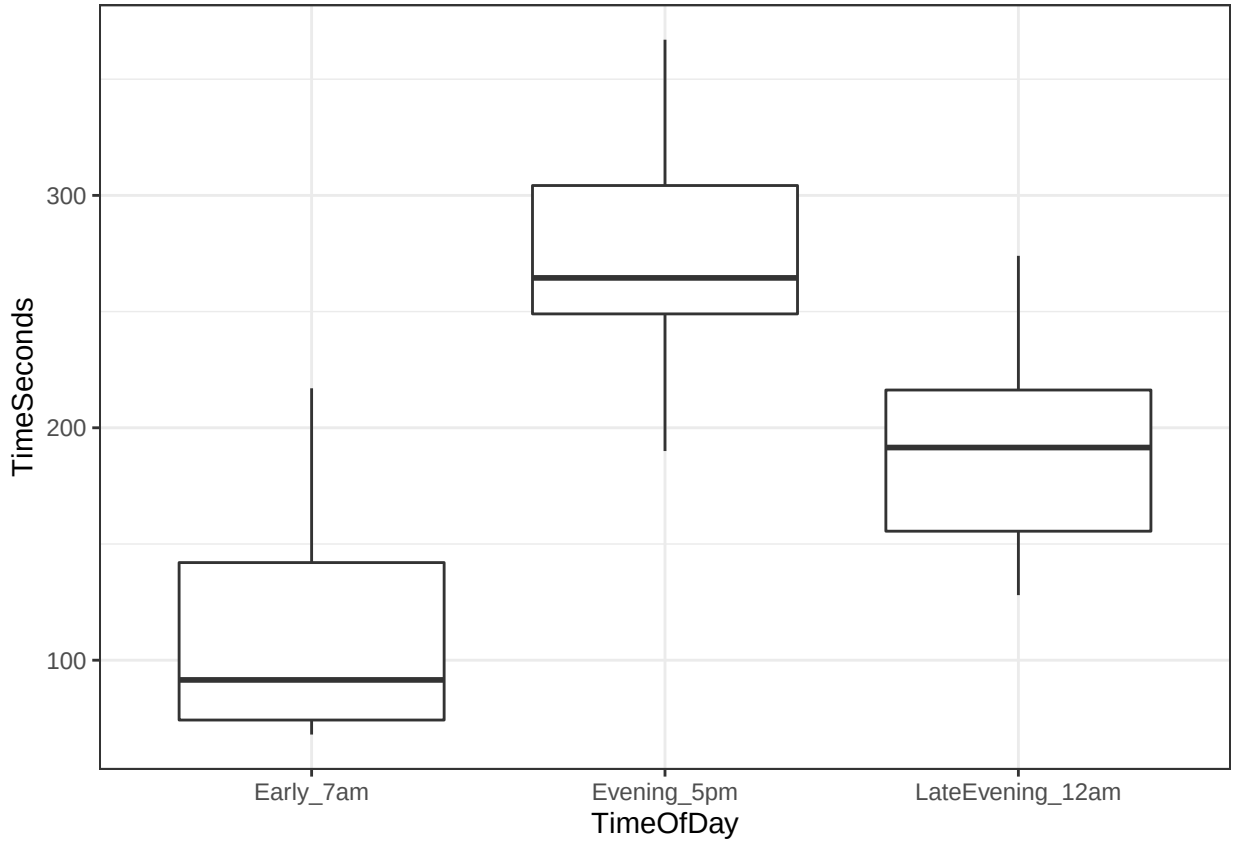
Reference sites:

- http://sci2s.ugr.es/keel/pdf/specific/articulo/35723.pdf
- http://brownmath.com/stat/anova1.htm

Example

Create Boxplots, check assumptions, perform ANOVA... if significant... perform Tukey posthoc comparisons.

```
TimeOfDay <- c(rep("Early_7am", 16), rep("Evening_5pm", 16), rep("LateEvening_12am", 16))
TimeSeconds <- c(68, 138, 75, 186, 68, 217, 93, 90, 71, 154, 166, 130, 72, 81, 76, 129, 299,
                 367, 331, 257, 260, 269, 252, 200, 296, 204, 190, 240, 350, 256, 282, 320,
                 216, 175, 274, 171, 187, 213, 139, 139, 226, 128, 236, 128, 217, 196, 201, 161)
DF <- data.frame(TimeOfDay, TimeSeconds)
rm("TimeOfDay", "TimeSeconds")
library(tidyverse)
ggplot(data = DF, aes(x = TimeOfDay, y = TimeSeconds)) +
  geom_boxplot() +
  theme_bw()
```




```
DF %>%
  group_by(TimeOfDay) %>%
  summarize(mean(TimeSeconds), sd(TimeSeconds))
```

```
# A tibble: 3 x 3
  TimeOfDay        `mean(TimeSeconds)` `sd(TimeSeconds)`
  <fct>                          <dbl>             <dbl>
1 Early_7am                       113.              47.7
2 Evening_5pm                     273.              52.2
3 LateEvening_12am                188.              42.3
```

```
anova(lm(TimeSeconds ~ TimeOfDay, data = DF))
```

```
Analysis of Variance Table
```

```
Response: TimeSeconds
          Df Sum Sq Mean Sq F value    Pr(>F)
TimeOfDay  2 204952  102476  45.325 1.652e-11 ***
Residuals 45 101742    2261
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod <- aov(TimeSeconds ~ TimeOfDay, data = DF))
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
TimeOfDay    2 204952  102476   45.33 1.65e-11 ***
Residuals   45 101742    2261
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(mod)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = TimeSeconds ~ TimeOfDay, data = DF)

$TimeOfDay
                                  diff        lwr       upr     p adj
Evening_5pm-Early_7am         159.9375  119.19361 200.68139 0.0000000
LateEvening_12am-Early_7am     74.5625   33.81861 115.30639 0.0001709
LateEvening_12am-Evening_5pm -85.3750 -126.11889 -44.63111 0.0000208
```

```
plot(TukeyHSD(mod))
```